Impact of the spatial context on human communication activity

Zolzaya Dashdorj[1] and Stanislav Sobolevsky[2]

[1]University of Trento and SKIL LAB - Telecom Italia and DKM - Fondazione Bruno Kessler, Italy Via Sommarive, 9 Povo, TN, Italy
dashdorj@disi.unitn.it

[2]Massachusetts Institute of Technology, MIT 77 Massachusetts Avenue Cambridge, MA, USA
stanly@mit.edu

Abstract

Technology development produces terabytes of data generated by human activity in space and time. This enormous amount of data often called big data becomes crucial for delivering new insights to decision makers. It contains behavioral information on different types of human activity influenced by many external factors such as geographic information and weather forecast. Early recognition and prediction of those human behaviors are of great importance in many societal applications like health-care, risk management and urban planning, etc. In this paper, we investigate relevant geographical areas based on their categories of human activities (i.e., working and shopping) which identified from geographic information (i.e., Openstreetmap). We use spectral clustering followed by k-means clustering algorithm based on TF/IDF cosine similarity metric. We evaluate the quality of those observed clusters with the use of silhouette coefficients which are estimated based on the similarities of the mobile communication activity temporal patterns. The area clusters are further used to explain typical or exceptional communication activities. We demonstrate the study using a real dataset containing 1 million Call Detailed Records. This type of analysis and its application are important for analyzing the dependency of human behaviors from the external factors and hidden relationships and unknown correlations and other useful information that can support decision-making.

Index terms— telecommunication dataset, human behavior, cell phone data records, activity recognition, knowledge management, clustering and classification

1 Introduction

Nowadays extensive penetration of digital technologies into everyday life creates vast amount of data related to different types of human activity. When available for the research purposes this creates an unprecedented opportunity for understanding human society directly from its digital traces. There is an impressive amount of papers leveraging such data for studying human behavior, including mobile phone records [1, 4, 9, 20, 21, 22], vehicle GPS traces [14, 24], social media posts [12, 13, 17] and bank card transactions [25, 26, 27]. And mobile phone data is among the most commonly used data sources from above. Such data allows us to understand human behaviors and social relationships by investigating the influence of context factors of social dynamics. Potential applications of this research are a context aware application systems, and "smart cities" applications that provide decision support for stakeholders in areas such as urban, transport planning, tourism and event analysis, emergency response, health improvement, community understanding, economic indicators and others. Current research [2, 3, 5, 8, 19] notes that with a pure CDR, it is possible to identify human behaviors, but results suffer from the heterogeneity, uncertainty and complexity of raw datasets and that the lack of qualitative content is included in the data itself that may be used to help to infer human behaviors. [6] identifies that Points of Interest (POIs) provide a good proxy for predicting the content of human activities in each area and thus for identifying the activities people are more likely to perform. This is much more effective if we combine mobile phone data records that can be more likely associated to human activities, for example, a person looking for some food if phone call is located in or close to a restaurant. In this paper, we concentrated on characterization and clustering of geographical areas based on the categories of human activity in order to identify relevant areas. The model proposed in [5], is used to extract top level human activities from geographical open data source. We use spectral clustering with eigengap heuristic followed by k-means clustering and intrinsic method to evaluate the quality of clusters. In each area cluster, we contextually enrich the mobile phone data records with the categories of human activities and then analyze and identify the standard or exceptional (divergent) type of the communication activity temporal patterns. This further opens a discussion

to understand various type of relationships between environment and human behavior. The paper is structured as follows Section 2 illustrates the related works and methodology is described in Section 4. We present and discuss the results in Section 5. Finally, we summarize the discussions in Section 6.

2 Related works

The clustering approaches (Han & Kamber [11]) such as k-means, k-medoids, and self organizing map group similar spatial objects into classes, and several other methods are also used to perform effective and efficient clustering, for instance, Calabrese et al [23] and also [10, 18] used eigengap heuristic for clustering. Phithakkitnukoon et al [19] identified area profile from POIs. Each area is connected to main activity considering the category of POIs, and activity patterns for each mobile user are studied to determine groups which have similar activity patterns. Noulas et al [16] proposed an approach for modeling and characterization of geographic areas based on a number of user check-ins and a set of 8 general (human) activity categories in Foursquare. Cosine similarity metric is used to measure a similarity of geographical areas, and spectral clustering algorithm that followed by k-means clustering is applied to identify a relevant area profile. The area profiles enables to understand groups of individuals who have similar activity patterns. Similar to this research idea, social networks[29] have been taken into account to discover activity patterns of individuals. Frias-Martinez et al [7] studied geolocated tweets to characterize urban landscapes using a complimentary source of land-use and landmark information. The author focused on determining the land-uses in a specific urban area based on tweeting patterns, and identification of POIs in high activity tweeted areas. Differently, Yuang et al [30] proposed to classify urban areas based on their mobility patterns by measuring the similarity between the time-series using Dynamic Time Warping (DTW) algorithm. Some of areas focus on understanding urban dynamics including dense area detection and their evolution over time [15, 28].

3 Data-source collection

We model contextual information of 4.6 million POIs in Trento, Italy and behavioral dataset of 1 million mobile phone data records (CDR).

Table 1: The categories of human activity belong to the POIs

Top level classes of activities	**Type of POIs**
eating	fast food, food court, restaurant, cafe
shopping	grocery, general stores
health medicine activity	hospital, pharmacy
entertainment activity	bar, casino, movie, theater
education activity	library, university school
transportation traveling	airplane, bus, car, train
outdoor activity	sightseeing, personal care, religious places
sporting activity	car racing, summer, winter sports
working activity	professional work place, industrial place
residential activity	guest house, hotel, hostel, residential building

3.1 Openstreetmap

For modeling the contextual description of geographical regions in Trento, we use a High level Representation of Behavior Model (HRBModel) [5, 6] which exploits a spatial grid (i.e., cell size is 50m x 50m) which populated the locations (i.e., cell) with POIs from open geographic information, Openstreetmap (OSM). The model generates human activity distribution map which enriched by the categories of human activity associated with a likelihood measure. For example, watching a football match on a stadium, eating in restaurant or hiking in forest. We collected in total 135,918 relevant POIs extracted from OSM. After cleaning and discarding irrelevant POIs (i.e., those that do not reflect relevant human activity), the number of POIs is reduced to 31,514 POIs. The total number of human activities are extracted up to 78,068 belongs to the POIs. The top-level classes of activity categories belong to the POIs are explained in Table 1.

3.2 Mobile phone data records

We collected CDR about outgoing logged calls for 2 months. The CDR is completely anonymized containing cell ID, time of day and duration in which a phone call is issued. Cell-ID is used for identification of some portion of a physical geographic area featured with a set of devices (antennas) that support the communication. Sometimes, cell coverage area is not precisely defined, and can be temporarily modified depending on the estimation of call traffic from/to this area. Usually the size of the coverage area is inversely proportional to the density of the population inhabiting the area. It is observed that, in presence of regular territory (i.e., flat with no mountains or other natural irregularities), the shape of cells can be approximated with convex polygons, otherwise the cell can be very irregular and possibly disconnected.

4 Methodology

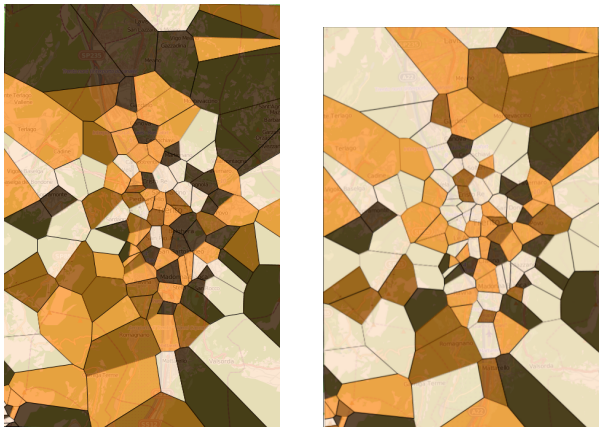
We present our approach for clustering algorithm to identify relevant areas in terms of geographical area profiles containing a set of the categories of human activity and use the observed clusters for analyzing mobile phone communication activities. Our approach is aimed at answering the following questions: What are the geographical area profiles defined by human activities in a city? How is the communication pattern affected by the profile of area activity? To do that, we first identify relevant area clusters based in terms of the category of activities and then analyze the communication activity patterns in those observed area clusters.

4.1 Geographical area clusters

We define a vector space model that contains a set of activity categories corresponding to geographical areas. The representation of geographic areas l_i within the territory of a city or large square area L. Each area l_i contains a vector of different top level activities derived from the POIs in such area that would be an input data-point for identifying the relevant areas by cluster algorithms. The area features are represented by a matrix $l_{i,j}$ containing the weight of the activity categories j in each area L_i. The relevance between the areas is identified by the cosine similarity metric by estimating the deviation of angles among area vectors. For example, the similarity between area l_1 and l_2 is as $\cos\theta_{1,2} = \frac{\mathbf{l_1} \cdot \mathbf{l_2}}{\|\mathbf{l_2}\| \|\mathbf{l_1}\|}$. Having the estimation of similarity between the areas, we can now create a similarity graph described as the weight matrix W and the degree matrix D is utilized by the spectral clustering algorithm which is the one of the most popular modern clustering methods and performs better than traditional clustering algorithms. The K-Nearest Neighbors of each data point are identified using cosine similarity metric, we create the adjacency matrix of the similarity graph and graph Laplacian $L = D - A$ (given by normalized graph Laplacian $L_n = D^{-1/2} L D^{-1/2}$). Based on eigengap heuristic, we identify the number of clusters to observe in our dataset as $k = argmax_i(\lambda_{i+1} - \lambda_i)$ where $\lambda_i \in \{l_1, l_2, l_3, .., l_n\}$ denotes the eigenvalues of L_n in the ascending order. Finally, we easily detect the effective clusters (area profiles) $C_1, C_2, C_3, ..., C_k$ from the first k eigenvectors identified by the k-means algorithms.

4.2 Behavioral pattern extraction in area clusters

We are interested in the contextualization of communication activity temporal patterns in relevant area clusters in order to determine standard or exceptional type of communication activities based on the communication activity variations in different time context. This could be done by overlapping between human activity distribution map and cell coverage map. However in this analysis, the cell coverage map is unavailable and the location of cell towers are given approximately. We decided to divide the study area into Voronoi polygons based on the spatial distribution of cell phone towers. The Figure 1(a) shows the Voronoi polygons for visualizing cell coverage map. We are then able to extract mobile communication activities in each polygon as a coverage area. For extracting mobile communication activities in observed area clusters, we need to associate each Voronoi polygon to the human activity distribution map. A given cell area p might be intersected with multiple areas l_i that represented as a list of areas with an intersection weight [0,1]. The intersection weight is estimated by the division of the areas size l_i and p, as $W(p, l_i) = \frac{S(l_i)}{S(p)}$. The number of calls per area is the number of calls in a given cell p at a certain time t divided by the number of intersecting areas N, taking the intersection weights into account, as $X(p, t, l_i) = \frac{X(p,t)}{N} \cdot w_i$.



(a) The density of communication activity distribution over typical day

(b) The density of communication activity distribution over Easter Sunday

Figure 1: The density of communication activity distribution is represented in the colors black belongs to high volume activity and white belongs low volume activity

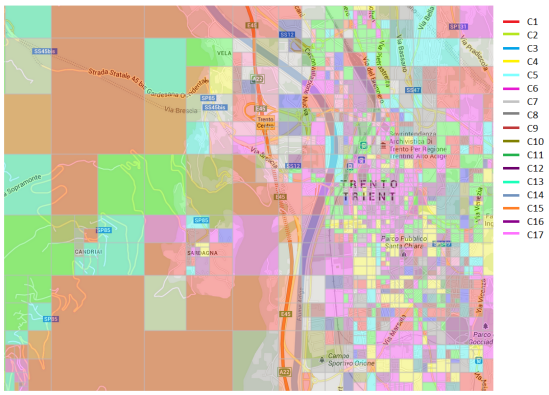


Figure 2: Geo-visualization of area classifications (k=17), Trento, Italy

We are now able to extract mobile communication activity patterns in different area clusters. In order to identify a normal (typical) type of communication activity temporal pattern, we exclude the communication activities over specific days when the public holidays or festivals occur, such as Liberation day, Palm Sunday and Easter Holiday (see the example during Easter holiday, Figure 1(b)) because the communication activities could have significant changes. To estimate the variation boundaries of the typical communication activities over different time context in a given area cluster we use a sigma approach in order to determine exceptional (divergent) communication activity temporal patterns in each area cluster, as $X'(c,t) = \mu_{X(c,t)} \pm \alpha \cdot \sigma_{X(c,t)}$.

5 Experimental Results and Discussion

We are concentrated in identifying relevant areas in which we show how semantics of human activities could be observed to interpret standard or exceptional communication activity temporal patterns. We observed 17 clusters (area profiles) according to activity vector of each area as shown in Figure 2. For each cluster, we show the weight vector of customer activities for each cluster as described in Figure 3. The central part of the city is clustered into C_6, C_2 followed by C_{12}, C_{16} where entertainment, residential, shopping, sporting and traveling by transport activities are highly distributed. We then extract the communication activity temporal patterns in order to see how communication behavior is affected by types of customer activity. The overall average communication activity density per day for each cluster varies depending on the category of the

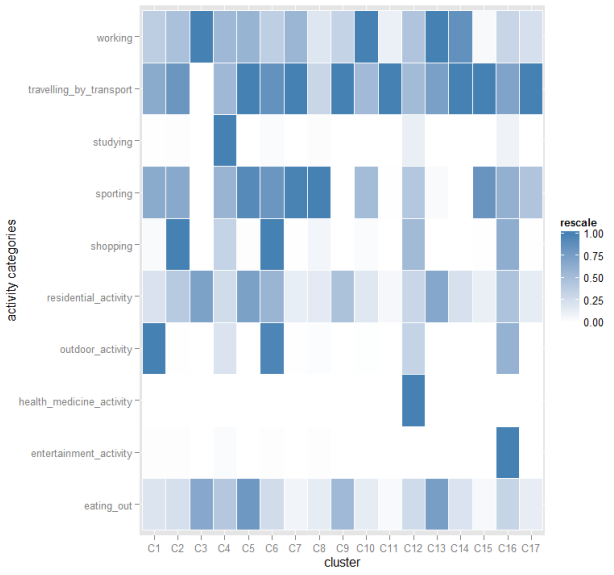


Figure 3: The scaled weight of activity categories of each cluster

customer activities as shown in Figure 4. C_1 is the most active cluster in terms of mobile phone activity communication. The weekly temporal communication activity variations per cluster are showing behaviors, similar to each other, see Figure 5.

Also there is clearly one purple cluster C_{11} (traveling by transport) which is more active over the weekdays compared to other clusters and less over the weekends and another one light-blue C_1 (outdoor activity), which shows the opposite pattern to C_{11}. This patterns are highlighted in the subplot of Figure 5 where the total percentage of weekend activity is reported. The different clusters patterns are different in terms of time context (e.g., hour of a day and day of a week).

The daily communication activity timeline over weekday vs weekend (saturday and sunday) are shown in figures 6(a), 6(b) and 6(c), respectively. They generally demonstrate quite similar pattern for different clusters. Based on the euclidean distance metric C_{11} is described as the most distinct cluster pattern to the average communication activity pattern over weekday, Saturday and Sunday.

To assess the accuracy of the cluster quality when the ground truth of a dataset is not available, we have to use an intrinsic method. We evaluate

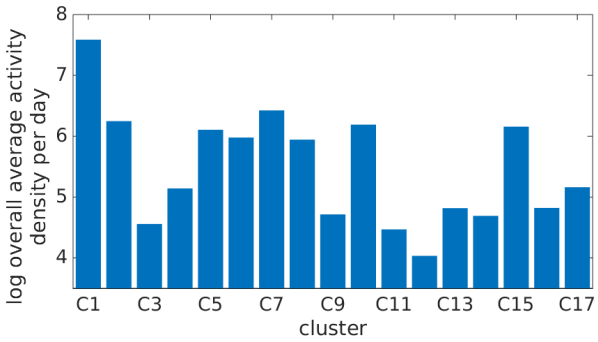


Figure 4: Overall average communication activity density (log scaled) per day with the actual values

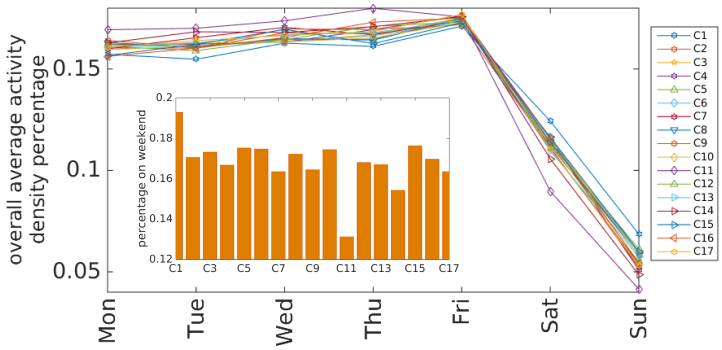
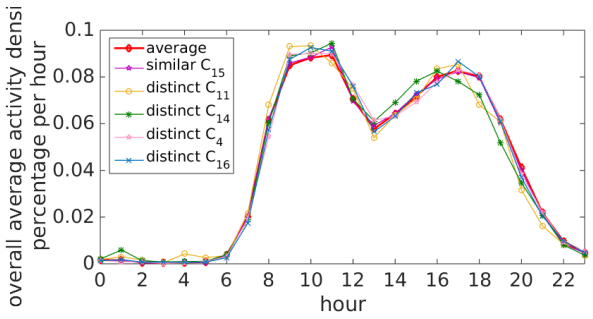
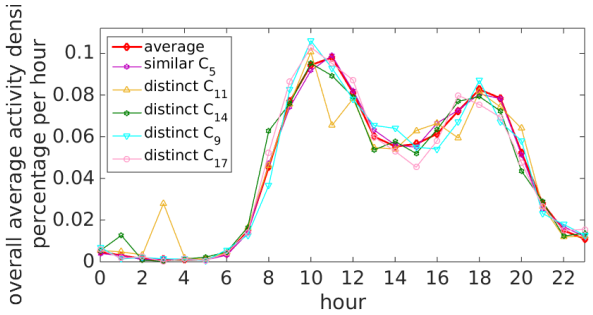


Figure 5: Communication activity temporal variations on the day of week and the subplot is about overall average communication activity density over weekend

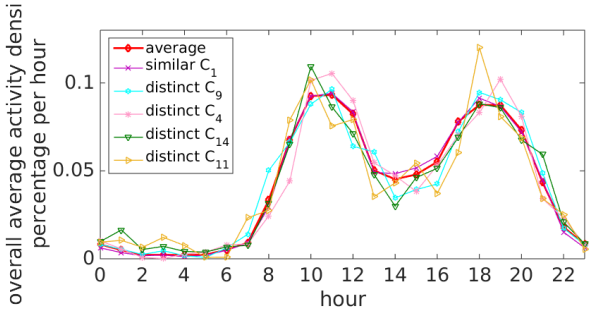
these clusters by examining how well the clusters are separated and how compact the clusters are, based on the similarity metric (silhouette coefficient) between objects in the dataset: $a(o) = \frac{\sum_{o' \in C_i, o \neq o'} dist(o,o')}{|C_i|-1}$, where $a(o)$ is the average distance between o and all other objects in the cluster to which o belongs. Similarly, $b(o)$ is the minimum average distance from o to all clusters to which o does not belong. Formally, suppose $o \in C_i (1 \leqslant i \leqslant k)$; then $b(o) = \min_{C_j : 1 \leq j \leq k, j \neq i} \left\{ \frac{\sum_{o' \in C_j} dist(o,o')}{|C_j|} \right\}$. The silhouette coefficient is between -1 and 1, estimated by $s(o) = \frac{b(o)-a(o)}{\max\{a(o),b(o)\}}$. The positive value reflects the more compactness of the cluster and well separated from other clusters. However, when the silhouette coefficient value is negative, the object in the con-



(a) Daily communication activity pattern per cluster over weekday. The cluster patterns are more diverse over weekend while the patterns are almost similar over weekdays



(b) Daily communication activity pattern per cluster over Saturday. In the pattern C_{11}, there is an activity peak at the 3:00am on Saturday which means in the late night of Friday, people take a transportation to go home



(c) Daily communication activity pattern per cluster over Sunday. In the pattern of cluster C_{14}, there is a peak every 1am that might be resulted from the traveling by transport and working activity

Figure 6: Typical communication activity temporal patterns over weekday vs weekend, the clusters C_{11} and C_{14} are the most distinct clusters compared to the average communication activity pattern over weekday, Saturday and Sunday

sidered same cluster is closer to the objects in another cluster. From the communication activity temporal patterns over the days of week, we estimated the silhouette coefficients of each area in all clusters based on the euclidean distance measure. This specific transport activity cluster C_{11} is estimated with the clustering quality of 67% where the silhouette coefficient of the objects in the cluster are positive. This means, the cluster is well separated from the other clusters and compact. This quality measure is increased to 77% when we estimate the coefficient from the communication activity temporal patterns over whole time periods. We picked it up for further analysis as we expect traveling to be largely affected by special events and also as this cluster is a particular one (in terms of deviations of the timeline from average), at the same time having quality measure.

We further investigate exceptional (divergent) type of temporal patterns over public events to determine how much the communication activity deviates from the variations of typical communication activities using the sigma approach. Figure 7 shows the changes of communication activity temporal pattern over Easter Sunday and Palm Sunday. The Easter Sunday has a great impact of human behaviors as there is an activity peak in the morning of the Easter Sunday.

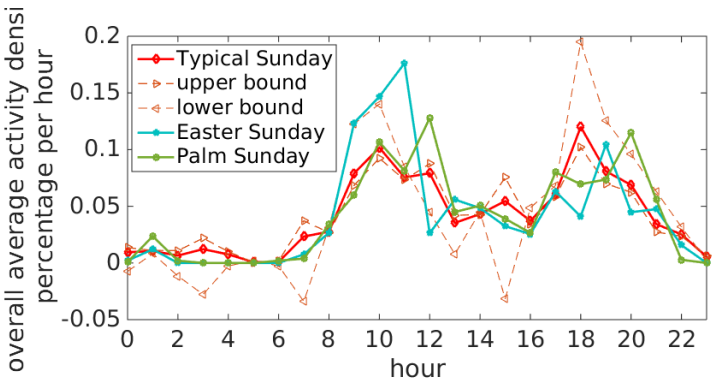


Figure 7: The daily temporal communication activity pattern of C_{11} over typical Sunday compared with the communication activity pattern over Easter Sunday and Palm Sunday. There is an activity peak in the morning of Easter Sunday which is relatively increased than the typical variations of the communication activity, $\alpha = 3$

6 Conclusion

In this paper, we proposed an approach to identify relevant area clusters in terms of the categories of human activity (i.e., working, shopping or entertainment areas). The area clusters are used to contextualize mobile communication activities temporal patterns with the categories of human activity. An intrinsic method is used to assess the clustering quality if the cluster is well separated from other clusters and compact. And it turns out that communication activity, namely its density and temporal variation - is largely affected by the context of human activity in the area. The transport activity cluster C_{11} is well classified with the clustering quality of 77%. With the use of those area clusters, we explain the typical or exceptional (divergent) type of mobile communication activities. In future works, we evaluate the approach in different cities and measure the relation between the other types of human activity and mobile communication activities. The result of the research work is potentially useful for more coherent classifications of human behaviors and better understanding the relationship between human behaviors and environmental factors and their dynamics in real-life social phenomena.

7 Acknowledgments

The authors would like to thank the Semantic Innovation Knowledge Lab - Telecom Italia for providing the mobile phone data records. We also would like to thank MIT SENSEable City Lab Consortium for supporting the research.

References

[1] A. Amini, K. Kung, C. Kang, S. Sobolevsky, and C. Ratti. The differing tribal and infrastructural influences on mobility in developing and industrialized regions. In *Proc. of 3rd int. conference on the analysis of mobile phone datasets*, 2013.

[2] J. P. Bagrow, D. Wang, and A.-L. Barabási. Collective response of human populations to large-scale emergencies. *CoRR*, abs/1106.0560, 2011.

[3] F. Calabrese, P. F. C., G. Di Lorenzo, L. Liu, and C. Ratti. The geography of taste: analyzing cell-phone mobility and social events. In *the Proc. of the 8th international conference on Pervasive Computing*, (Pervasive'10), pages 22–37, Berlin, Heidelberg, 2010. Springer-Verlag.

[4] F. Calabrese and C. Ratti. Real time rome. *Networks and Communication studies*, 20(3-4):247–258, 2006.

[5] Z. Dashdorj, L. Serafini, F. Antonelli, and R. Larcher. Semantic enrichment of mobile phone data records. In *12th International Conference on Mobile and Ubiquitous Multimedia, MUM '13, Luleå, Sweden - December 02 - 05, 2013*, page 35. ACM, 2013.

[6] Z. Dashdorj, S. Sobolevsky, L. Serafini, F. Antonelli, and C. Ratti. Semantic enrichment of mobile phone data records using background knowledge. *arXiv preprint arXiv:1504.05895*, 2015.

[7] V. Frías-Martínez, V. Soto, H. Hohwald, and E. Frías-Martínez. Characterizing urban landscapes using geolocated tweets. In *SocialCom/PASSAT*, pages 239–248, 2012.

[8] B. Furletti, L. Gabrielli, C. Renso, and S. Rinzivillo. Identifying users profiles from mobile calls habits. In *Proceedings of the ACM SIGKDD International Workshop on Urban Computing*, UrbComp '12, pages 17–24, New York, NY, USA, 2012. ACM.

[9] F. Girardin, F. Calabrese, F. D. Fiore, C. Ratti, and J. Blat. Digital footprinting: Uncovering tourists with user-generated content. *Pervasive Computing, IEEE*, 7(4):36–43, 2008.

[10] S. Grauwin, S. Sobolevsky, S. Moritz, I. Gódor, and C. Ratti. Towards a comparative science of cities: using mobile traffic records in New York, London and Hong Kong. *arXiv preprint arXiv:1406.4400*, 2014.

[11] J. Han, M. Kamber, and A. K. H. Tung. *Spatial Clustering Methods in Data Mining: A Survey*. Taylor and Francis, 2001.

[12] B. Hawelka, I. Sitko, E. Beinat, S. Sobolevsky, P. Kazakopoulos, and C. Ratti. Geo-located twitter as proxy for global mobility pattern. *Cartography and Geographic Information Science*, pages 1–12, 2014.

[13] A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 56–65. ACM, 2007.

[14] C. Kang, S. Sobolevsky, Y. Liu, and C. Ratti. Exploring human movements in singapore: a comparative analysis based on mobile phone and taxicab

usages. In *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing*, page 1. ACM, 2013.

[15] J. Ni and C. V. Ravishankar. Pointwise-dense region queries in spatio-temporal databases. In R. Chirkova, A. Dogac, M. T. zsu, and T. K. Sellis, editors, *ICDE*, pages 1066–1075. IEEE, 2007.

[16] A. Noulas, S. Scellato, C. Mascolo, and M. Pontil. Exploiting semantic annotations for clustering geographic areas and users in location-based social networks. In *The Social Mobile Web*, 2011.

[17] S. Paldino, I. Bojic, S. Sobolevsky, C. Ratti, and M. C. González. Urban magnetism through the lens of geo-tagged photography. *EPJ Data Science*, 4(1):1–17, 2015.

[18] T. Pei, S. Sobolevsky, C. Ratti, S.-L. Shaw, T. Li, and C. Zhou. A new insight into land use classification based on aggregated mobile phone data. *International Journal of Geographical Information Science*, 28(9):1988–2007, 2014.

[19] S. Phithakkitnukoon, T. Horanont, G. Di Lorenzo, R. Shibasaki, and C. Ratti. Activity-aware map: Identifying human daily activity pattern using mobile phone data. In *the Proc. of the 1st Intl. Conf. Human Behavior Understanding*, pages 14–25, 2010.

[20] D. Quercia, N. Lathia, F. Calabrese, G. Di Lorenzo, and J. Crowcroft. Recommending social events from mobile phone location data. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pages 971 –976, 2010.

[21] C. Ratti, S. Sobolevsky, F. Calabrese, C. Andris, J. Reades, M. Martino, R. Claxton, and S. H. Strogatz. Redrawing the map of Great Britain from a network of human interactions. *PLoS One*, 5(12):e14248, 2010.

[22] C. Ratti, S. Williams, D. Frenchman, and R. Pulselli. Mobile landscapes: Using location data from cell phones for urban analysis. *Environment and planning B*, 33(5):727, 2006.

[23] J. Reades, F. Calabrese, and C. Ratti. Eigenplaces: analysing cities using the space time structure of the mobile phone network. *Environment and Planning B: Planning and Design*, 36(5):824–836, 2009.

[24] P. Santi, G. Resta, M. Szell, S. Sobolevsky, S. H. Strogatz, and C. Ratti. Quantifying the benefits of vehicle pooling with shareability networks. *Proceedings of the National Academy of Sciences*, 111(37):13290–13294, 2014.

[25] S. Sobolevsky, I. Bojic, A. Belyi, I. Sitko, B. Hawelka, J. M. Arias, and C. Ratti. Scaling of city attractiveness for foreign visitors through big data of human economical and social media activity. *arXiv preprint arXiv:1504.06003*, 2015.

[26] S. Sobolevsky, I. Sitko, R. T. D. Combes, B. Hawelka, J. M. Arias, and C. Ratti. Money on the move: Big data of bank card transactions as the new proxy for human mobility patterns and regional delineation. the case of residents and foreign visitors in spain. In *Big Data (BigData Congress), 2014 IEEE International Congress on*, pages 136–143. IEEE, 2014.

[27] S. Sobolevsky, I. Sitko, S. Grauwin, R. T. d. Combes, B. Hawelka, J. M. Arias, and C. Ratti. Mining urban performance: Scale-independent classification of cities based on individual economic transactions. *arXiv preprint arXiv:1405.4301*, 2014.

[28] M. R. Vieira, V. Frias-Martinez, N. Oliver, and E. Frias-Martinez. Characterizing dense urban areas from mobile phone-call data: Discovery and social dynamics. In *Proceedings of the 2010 IEEE Second International Conference on Social Computing*, SOCIALCOM '10, pages 241–248, Washington, DC, USA, 2010. IEEE Computer Society.

[29] S. Wakamiya, R. Lee, and K. Sumiya. Urban area characterization based on semantics of crowd activities in twitter. In *Proceedings of the 4th international conference on GeoSpatial semantics*, GeoS'11, pages 108–123, Berlin, Heidelberg, 2011. Springer-Verlag.

[30] Y. Yuan and M. Raubal. Extracting dynamic urban mobility patterns from mobile phone data. In N. Xiao, M.-P. Kwan, M. F. Goodchild, and S. Shekhar, editors, *GIScience*, volume 7478 of *Lecture Notes in Computer Science*, pages 354–367. Springer, 2012.